#Principal Component Analysis.

Considering a dataset with n samples and d features, representing it in matrix format,

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in R^{n \times d}$$

each row → observation
each column → feature.

for PCA, we need to find new axis such that,

maximum variance is captured,
axis are orthogonal and dimentionality can be reduced

computing mean of each feature,

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \quad ; \text{centered data} : X_c = X - \mu$$

$$Cov(x1, x2) = \frac{\sum_{i=1}^{n} (x1_i - \bar{x1})(x2_i - \bar{x2})}{n}$$

$$X_c = \begin{bmatrix} x_1 - \mu \\ x_2 - \mu \\ \vdots \\ x_n - \mu \end{bmatrix}$$

defining covariance matrix

$$C = \frac{1}{n} X_c^T X_c \qquad C \in R^{d \times d}$$

$$C_{ij} = Cov(feature_i, feature_j)$$

(All diagonal element is variance, and off-diagonal elements are co-variance)

We want to find a unit vector

$w \in \mathbb{R}^d$, such that projection has maximum variance

$w^T\left(\frac{1}{n}\sum x_i\right) = 0$

Projection of data onto w

$x_i \in \mathbb{R}^d$

$$z_i = w^T x_i$$

$z_i = w^T x_i$

(Converts d-dimensional data into scalar)

Variance of projected data,

$$var(z) = \frac{1}{n}\sum (w^T x_i)^2$$

$var(z) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2$

$\bar{z} = \frac{1}{n}\sum z_i = \frac{1}{n}\sum w^T x_i$

Matrix form $var(z) = \frac{1}{n}\|X_c w\|^2 = \frac{1}{n}(X_c w)^T(X_c w)$

$$= \frac{1}{n} w^T X_c^T X_c w = w^T C w$$

$\because C = \frac{1}{n} X_c^T X_c$

Optimization problem,

$$\max_w \; w^T C w \quad \text{subject to } w^T w = 1$$

(unit vector constraint)

Solving using Lagrange multipliers, → constrained optimization

$$L(w, \lambda) = w^T C w - \lambda(w^T w - 1)$$

objective; constraint enforcement; λ: eigen value; w: eigen vector

$$\frac{\partial L}{\partial w} = 2Cw - 2\lambda w = 0 \Rightarrow Cw = \lambda w$$

This is the eigen value eqn.

$\frac{\partial}{\partial w} w^T A w = 2Aw$

Given A is symm

variance along direction w is $Var = \lambda$.

$\frac{\partial}{\partial w} w^T w$

$w^T w = \sum_{i=1}^{d} w_i^2$

$\frac{\partial}{\partial w}\sum w_i^2 = 2w$.

$\therefore$ Largest eigen value → max variance direction

Corresponding eigen vector → first PCA.

for all eigen vectors $C = V \Lambda V^T$, where $V = [v_1, v_2, \ldots, v_d]$

eigen values sorted $\lambda \geq \lambda_2 \geq \lambda_3 \ldots$

The above equation comes from eigen value decomposition (spectral decomposition) of a symmetric matrix. The Covariance matrix C in PCA is always symmetric.

$\therefore PC1 = V_1, \; PC2 = V_2, \; PC_3 = V_3$

Project data, $Z = X_c V$

if reducing dimension to k:

$Z = X_c V_k$

$Z \in \mathbb{R}^{n \times k}$.

Q1. Given the data in the table, reduce the dimension from 2 to 1 using principal Component Analysis (PCA).

Soln

Feature	Example 1	Example 2	Copy 3	Copy 4
x_1	4	8	13	7
x_2	11	4	5	14

computing mean,

$\bar{x}_1 = \frac{1}{4}(4+8+13+7) = 8;$

$\bar{x}_2 = \frac{1}{4}(11+4+5+14) = 8.5$

computing Covariance matrix,

$$Cov(x_1, x_2) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{1k}-\bar{x})^2$$

$$= \frac{1}{3}\left[(4-8)^2+(8-8)^2+(13-8)^2+(7-8)^2\right]$$

$$= 14$$

$$Cov(x_1, x_2) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{1k}-\bar{x})(x_{2k}-\bar{x}_2)$$

$$= \frac{1}{3}\left[(4-8)(11-8.5)+(8-8)(4-8.5)+(13-8)(5-8.5)+(7-8)(14-8.5)\right]$$

$$= -11$$

$$Cov(x_2, x_1) = -11$$

$$Cov(x_2, x_2) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{2k}-\bar{x}_2)^2 = \frac{1}{3}\left[(11-8.5)^2+(4-8.5)^2+(5-8.5)^2+(14-8.5)^2\right] = 23$$

∴ Covariance Matrix

$$S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Eigen values of Covariance matrix,

$$\det(S-\lambda I)=0 \Rightarrow \begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (14-\lambda)(23-\lambda) - (-11)(-11) = \lambda^2 - 37\lambda + 201$$

Solving the characteristics equation,

$$\lambda = \frac{1}{2}\left(37 \pm \sqrt{365}\right)$$

$$= 30.38 \;, 6.615$$

$$\therefore \lambda_1 = 30.38 \quad ; \lambda_2 = 6.62$$

Satisfying the following equation,

$$(S-\lambda I)X = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} (14-\lambda_1)u_1 - 11u_2 \\ -11u_1 + (23-\lambda_1)u_2 \end{bmatrix}$$

$$(14-\lambda_1)u_1 - 11u_2 = 0$$

$$-11u_1 + (23-\lambda_1)u_2 = 0$$

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda_1} = t,$$

$u_1 = 11t$, $u_2 = (14-\lambda_1)t$, t is any real no

taking $t=1$; $U_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix}$

finding eigen vectors, we compute length of x_1

$$\|u_1\| = \sqrt{(11)^2 + (14-\lambda_1)^2}$$

$$= 19.73$$

∴ unit vector corresponding to λ_1 is given as,

$$e_1 = \begin{bmatrix} 11/\|u_1\| \\ (14-\lambda_1)/\|u_1\| \end{bmatrix} = \begin{bmatrix} 11/19.7348 \\ (14-30.38)/19.73 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

Carrying out similar computations,

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Computing first principal component, let $\begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix}$ be the k^{th} sample in the above table, the first principal component is given by,

$$e_1^T \begin{bmatrix} x_{1k}-\bar{x}_1 \\ x_{2k}-\bar{x}_2 \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} x_{1k}-\bar{x}_1 \\ x_{2k}-\bar{x}_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

and similarly, $e_2 \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$

$$= 4.3035$$

Computing for other k's we get the first PCA s,

-4.3052 $\quad$ 3.7361 $\quad$ 5.6928 $\quad$ -5.1238.

Linear Discriminant Analysis

Linear Discriminant Analysis is a supervised dimensionality reduction technique and classification technique. Its goal is to, project high dimensional data into lower-dimensional space, while maximizing class-separability.

Considering a dataset with 'n' samples, each sample has 'd' dimensional features. Number of classes = 'c'.

Let ~~$x_i \in \mathbb{R}^d$~~ $x_i \in \mathbb{R}^d$

and class labels, $y_i \in \{1, 2, 3, \cdots, c\}$

We want to have a projection vector, s.t. $w \in \mathbb{R}^d$ so that the projected value becomes, $z = w^T x$.

This converts d-dimensional data into 1D.

After projection, same class points should be close together and different class points should be far apart.

Computing mean of each class (considering 2 classes).

$$\mu_1 = \frac{1}{n_1} \sum_{x \in c_1} x_i \quad \text{and} \quad \mu_2 = \frac{1}{n_2} \sum_{x_i \in c_2} x_i$$

[vectors of dimension d×1]

Projecting mean into w,

$$m_1 = w^T \mu_1 \; ; \; m_2 = w^T \mu_2$$

$$|m_1 - m_2| = |w^T \mu_1 - w^T \mu_2| = |w^T (\mu_1 - \mu_2)|$$

Distance bet' projected mean

We want $|m_1 - m_2|$ to be large.

Computing variance within each class,

$$s_1^2 = \sum_{x_i \in C_1} (w^T x_i - w^T \mu_1)^2 = \sum_{x_i \in C_1} (w^T (x_i - \mu))^2$$

$$= w^T \left(\sum_{x_i \in C_1} (x_i - \mu_1)(x_i - \mu_1)^T \right) w$$

Define: $$S_1 = \sum_{x_i \in C_1} (x_i - \mu_1)(x_i - \mu_1)^T$$

$$\therefore s_1^2 = w^T S_1 w$$

Similarly, $s_2^2 = w^T S_2 w$

Defining within class scatter matrix,

$$S_w = S_1 + S_2$$

Projected within-class variance

$$s_w^2 = w^T S_w w$$

Define between class scatter,

Distance between projected means $(m_1 - m_2)^2 = (w^T (\mu_1 - \mu_2))^2$

$$= w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w$$

$$S_B = (\mu_1 - \mu_2)(\mu - \mu_2)^T$$

So, $w^T S_B w$

Defining the objective function,

$$J(w) = \frac{w^T S_B w}{w^T S_w w} \rightarrow \text{Maximize the objective function.}$$

$$J(w) = \frac{w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w}{w^T S_w w}.$$

This is maximized when, $w = S_w^{-1}(\mu_1 - \mu_2)$.

The optimal projection direction is, direction between means adjusted by variance. If variance is large, the importance should be reduced and vice-versa.

After finding w,

$$z = w^T x$$

classification can be done using a threshold.